Abstract of the Disclosure

A system and method for generating photo-realistic talking-head animation from a text input utilizes an audio-visual unit selection process. The lip-synchronization is obtained by optimally selecting and concatenating variable-length video units of the mouth area. The unit selection process utilizes the acoustic data to determine the target costs for the candidate images and utilizes the visual data to determine the concatenation costs. The image database is prepared in a hierarchical fashion, including high-level features (such as a full 3D modeling of the head, geometric size and position of elements) and pixel-based, low-level features (such as a PCA-based metric for labeling the various feature bitmaps).